

Simulation-based Bayesian analysis for multiple changepoints

Jason Wyse and Nial Friel

University College Dublin, Belfield, Dublin 4, Ireland

`jason.wyse@ucd.ie`, `nial.friel@ucd.ie`

November, 2010

Abstract: This paper presents a Markov chain Monte Carlo method to generate approximate posterior samples in retrospective multiple changepoint problems where the number of changes is not known in advance. The method uses conjugate models whereby the marginal likelihood for the data between consecutive changepoints is tractable. Inclusion of hyperpriors gives a near automatic algorithm providing a robust alternative to popular filtering recursions approaches in cases which may be sensitive to prior information. Three real examples are used to demonstrate the proposed approach.

Keywords: Bayes factor; changepoint; marginal likelihood; model search.

1 Introduction

The range of applications of changepoint models is evident from the substantial volume of literature devoted to this problem in the econometrics, signal processing and bioinformatics literatures. A process generating data can often undergo changes over time such that one model will not be appropriate for all time periods. Here “time” refers to some natural sequential indexing of the data. Some examples are occurrences of coal mining disasters during the 18th and 19th century (Raftery & Akman 1986), DNA or protein composition analysis over base number (Liu & Lawrence 1999) and winning streaks in sports (Yang 2004).

Markov chain Monte Carlo (MCMC) techniques can be used to estimate models with a fixed number of changepoints. When the number of changepoints is unknown, inference is more challenging. Chib (1998) estimates a collection of changepoint models and compares these using Bayes factors estimated from the MCMC output. Green (1995) uses reversible jump MCMC (RJMCMC) to explore the number of changepoints in the coal mining disaster data. RJMCMC allows moves between models which satisfy detailed balance.

The use of alternatives to MCMC has grown in this area in recent years. Fearnhead (2006) uses filtering recursions to derive the posterior distribution of changepoints. This can be done for both a known and unknown number of changepoints. An advantage of this approach is that one can draw independent samples from the posterior. MCMC can only

do this approximately at best. Extension to online analysis of changepoint models is also possible (Fearnhead & Liu 2007). However methods based on filtering recursions rely on strong prior information in most cases. This paper aims to offer an efficient MCMC alternative which can overcome strong reliance on prior assumptions as encountered in recursive computing approaches. The class of models considered is similar to Fearnhead (2006). For this reason it is possible that this could be used to give useful starting values for an analysis using filtering recursions.

Qualitatively, the work in this paper is similar in some aspects to work by Lavielle & Lebarbier (2001) and Punskeya et al. (2002) in terms of the class of models considered. The sampling aspect of the approach bears similarities to the samplers of Lavielle & Lebarbier (2001) and Girón et al. (2007). This paper extends these works to a broader range of data models and proposes a more efficient way of sampling changepoints. An aim is also to highlight possible shortcomings of alternatives to MCMC and how these could be overcome by using simulation approaches to inform choices for recursive computing approaches.

The remainder of the paper is organised as follows. In Section 2 the type of changepoint model under consideration is presented. Section 3 reviews the reversible jump approach to changepoint estimation and discusses how this can be simplified into a fixed dimensional sampling scheme. Section 4 gives the moves to sample from the simpler fixed dimensional posterior. Prior specification is discussed in Section 5, and Section 6 reviews the filtering recursion approach to generating samples of changepoints. Performance of the sampler is validated by analyzing the coal mining disasters data in Section 7, while Sections 8 and 9 compare qualitative aspects of the simulation based sampler approach and filtering recursions approach using two real data examples. A brief discussion concludes the article.

2 Changepoint models

Consider the data $y_{1:n} = (y_1, \dots, y_n)$ which is time ordered. Here y_i is observed before y_j if $i < j$. Time in this context can refer to any natural ordering of the data as it is observed. A changepoint occurs at time t if y_1, \dots, y_t are generated differently to y_{t+1}, \dots, y_n . Referring to $y_{s:r}$ ($s < r$) as a segment, this says that the segments $y_{1:t}$ and $y_{t+1:n}$ are heterogeneous between but homogeneous within. Parametric changepoint models assign a different parameter for each segment to account for this heterogeneity.

This paper considers multiple changepoints which will be denoted τ_1, \dots, τ_k . These split the data into $k+1$ segments. The likelihood for segment j has parameter θ_j . Conditional on a segmentation, the data within each segment is assumed independent. It is also assumed that the regime parameters θ_j are independent. The likelihood of the segmentation $\tau = (\tau_1, \dots, \tau_k)$ is

$$\prod_{j=1}^{k+1} \prod_{i=\tau_{j-1}+1}^{\tau_j} \pi(y_i|\theta_j)$$

where for convenience $\tau_0 = 0, \tau_{k+1} = n$. Instead of using τ , segmentations can be labelled with the binary latent vector $z = (z_1, \dots, z_n)$ with $z_t = 1$ indicating a changepoint at time t and $z_n = 0$. Independent priors are assumed for each member of $\theta = (\theta_1, \dots, \theta_{k+1})$ with hyperparameter γ and there is a prior for the changepoints with hyperparameter ξ , given by

$\pi(z|k, \xi)$. The posterior may be written

$$\begin{aligned}\pi(z, \theta|y, k, \xi, \gamma) &\propto \pi(z|k, \xi)\pi(\theta|k, \gamma)\pi(y|\theta, z, k) \\ &= \pi(z|\xi) \prod_{j=1}^{k+1} \pi(\theta_j|\gamma) \prod_{i=\tau_{j-1}+1}^{\tau_j} \pi(y_i|\theta_j)\end{aligned}$$

where the dependence on the number of changepoints, k , is made explicit. A prior $\pi(k)$ may be introduced so that the posterior of interest is the joint posterior of (k, z, θ) ,

$$\pi(k, z, \theta|y, \xi, \gamma) \propto \pi(k)\pi(z, \theta|y, k, \xi, \gamma). \quad (1)$$

This is a hierarchical changepoint model similar to that used in Green (1995).

3 Collapsing changepoint models

It is possible to construct a MCMC scheme to sample the posterior of (1) using RJMCMC (Green 1995). The sampler will explore the product space support of this posterior:

$$\mathcal{X} = \prod_k \{k\} \times \{\mathcal{Z}_k, \Theta_k|k\}$$

where \mathcal{Z}_k, Θ_k are respectively the sample spaces of z and θ conditional on k changepoints. A switch in the number of changepoints in the model can be made by a RJ move switching between support subspaces. For the purposes of illustration a straightforward move of this type is now discussed. When proposing a switch from k to $k+1$ changepoints one possibility is to generate a random variable $u \in \mathbb{R}^d$ and form a bijection $f : \Theta_k \times \mathbb{R}^d \rightarrow \Theta_{k+1}$ where d is the dimension of a single θ_j . This bijection gives the parameters for the proposed $k+1$ changepoint model as a function of those for the k changepoint model; $\theta' = (\theta'_1, \dots, \theta'_{k+2}) = f(\theta_1, \dots, \theta_{k+1}, u)$. The proposed switch in model is then accepted with probability $\min(1, R)$ where

$$R = \frac{\pi(k+1, z', \theta'|y, \xi, \gamma)}{\pi(k, z, \theta|y, \xi, \gamma)} \frac{P(k+1, k)}{P(k, k+1)} \frac{1}{q(u|\theta)} \left| \frac{\partial(\theta')}{\partial(\theta, u)} \right|.$$

In the expression for R , $P(\cdot, \cdot)$ denotes the proposal probability for transitions between different numbers of changepoints, and $q(\cdot|\theta)$ is the proposal density of u . The last term on the right is a Jacobian term for the bijection f . The reverse move in switching from $k+1$ to k changepoints is accepted with probability $\min(1, R^{-1})$. More elaborate moves between support subspaces are possible which propose changes to the model of more than one dimension or involve stochastic moves in both directions.

The key questions in a changepoint analysis are usually; how many changepoints are there and where are the changepoints? The segment parameters θ can be viewed as a nuisance parameter in this regard. Choosing conjugate priors for the θ_j allows these to be collapsed in the model

$$\begin{aligned}\pi(k, z|y, \xi, \gamma) &\propto \pi(k)\pi(z|k, \xi) \prod_{j=1}^{k+1} \int \pi(\theta_j|\gamma) \prod_{i=\tau_{j-1}+1}^{\tau_j} \pi(y_i|\theta_j) d\theta_j \\ &= \pi(k)\pi(z|k, \xi) \prod_{j=1}^{k+1} \pi(y_{\tau_{j-1}+1:\tau_j}|\gamma),\end{aligned} \quad (2)$$

where $\pi(y_{\tau_{j-1}+1:\tau_j}|\gamma)$ is the marginal likelihood of the data segment $y_{\tau_{j-1}+1:\tau_j}$ and is assumed to be available in closed form due to the conjugacy. The support of this posterior is

$$\mathcal{Y} = \prod_k \{k\} \times \{\mathcal{Z}_k|k\}$$

and a switch from k to $k+1$ changepoints does not require the design of a bijective function between support subspaces. The proposed switch in model is now accepted with Metropolis-Hastings probability $\min(1, A)$ where

$$A = \frac{\pi(k+1, z'|y, \xi, \gamma)}{\pi(k, z|y, \xi, \gamma)} \frac{P(k+1, k)}{P(k, k+1)}. \quad (3)$$

This idea of collapsing has been used previously in Panskaya et al. (2002) and Lavielle & Lebarbier (2001) for Gaussian data models.

It can be seen that the first term on the right hand side of the acceptance ratio (3) is the Bayes factor for a model with $k+1$ changepoints at positions z' versus a model with k changepoints at positions z , assuming all models are equally likely, *a priori*. Noting this, it becomes apparent that sampling k and z is equivalent to a model search over large model space. If there can be at most \bar{k} changepoints, then the dimension of this space is $\sum_{k=0}^{\bar{k}} \binom{n-1}{k}$. So searching for up to 5 changepoints in a dataset of length 200 corresponds to a dimension $\sim 2.5 \times 10^9$. In the next section an MCMC scheme to search over these large model spaces, that is, sample from the posterior (2), is proposed.

4 Sampling changepoints

The MCMC scheme to generate samples of changepoints from the posterior (2) consists of three possible moves: add a changepoint; delete a changepoint; move a changepoint. Each sweep consists of the following;

- i. Choose to add or delete a changepoint with probabilities a_k and $d_k = 1 - a_k$ respectively. Clearly $a_{\bar{k}} = d_0 = 0$.
- ii. Select a changepoint and propose to move it to a position in the range of its closest neighbouring changepoints.

Add or delete a changepoint

This move has been discussed in Section 3 but more details are given here. Suppose there is currently k changepoints at positions z . Let z correspond to changepoints at τ_1, \dots, τ_k . Randomly select one of the $n - k - 1$ points where there could be a changepoint i.e. a $t < n$ with $z_t = 0$. Say this is currently in segment j given by $y_{\tau_{j-1}+1:\tau_j}$. Relabel the proposed changepoints in z' as $\tau'_1, \dots, \tau'_{k+1}$ with $\tau'_j = t$. Cancellation of marginal likelihood terms then implies that

$$\frac{\pi(k+1, z'|y, \xi, \gamma)}{\pi(k, z|y, \xi, \gamma)} = \frac{\pi(k+1)}{\pi(k)} \frac{\pi(z'|k+1, \xi)}{\pi(z|k, \xi)} \frac{\pi(y_{\tau'_{j-1}+1:\tau'_j}|\gamma)\pi(y_{\tau'_j+1:\tau'_{j+1}}|\gamma)}{\pi(y_{\tau_{j-1}+1:\tau_j}|\gamma)}$$

so calculation of A in (3) only requires at most three marginal likelihood values. Conversely, for the delete move, one of the $k+1$ changepoints in z' is chosen at random and the calculation of the acceptance probability involves

$$\frac{\pi(k, z|y, \xi, \gamma)}{\pi(k+1, z'|y, \xi, \gamma)} = \frac{\pi(k)}{\pi(k+1)} \frac{\pi(z|k, \xi)}{\pi(z'|k+1, \xi)} \frac{\pi(y_{\tau_{j-1}+1:\tau_j}|\gamma)}{\pi(y_{\tau'_{j-1}+1:\tau'_j}|\gamma)\pi(y_{\tau'_j+1:\tau'_{j+1}}|\gamma)}.$$

Finally, the proposal one step transition probabilities for the number of changepoints will be $P(k, k+1) = a_k/(n-k-1)$ and $P(k+1, k) = d_{k+1}/(k+1)$, so that A (3) can be computed. The acceptance probability for the add move is then $\min(1, A)$ and the delete move is accepted with probability $\min(1, A^{-1})$.

Move a changepoint

Gibbs update: Given the model assumption that the marginal likelihood for any segment is available in closed form, it is possible to update the position of any changepoint from its full conditional. Suppose τ_j is being updated. Then the conditional probability that $\tau_j = t$, $\tau_{j-1} < t < \tau_{j+1}$ is proportional to

$$\pi(z'_{(t)}|k)\pi(y_{\tau_{j-1}+1:t}|\gamma)\pi(y_{t+1:\tau_{j+1}}|\gamma)$$

where $z'_{(t)}$ corresponds to changepoints $\tau_1, \dots, \tau_{j-1}, t, \tau_{j+1}, \dots, \tau_k$. The effort required for the Gibbs update is $O(\tau_{j+1} - \tau_{j-1})$ and so may be computationally expensive for large datasets with changepoints far apart, or datasets with many changepoints. In this situation a local random walk update may be preferred.

Local random walk update: t is drawn uniformly from the integers $\max(\tau_j - l, \tau_{j-1} + 1), \dots, \min(\tau_j + l, \tau_{j+1} - 1)$ where l specifies the locality of the proposed move. The move is accepted with probability $\min(1, B)$ where

$$B = \frac{\pi(y_{\tau_{j-1}+1:t}|\gamma)\pi(y_{t+1:\tau_{j+1}}|\gamma)}{\pi(y_{\tau_{j-1}+1:\tau_j}|\gamma)\pi(y_{\tau_j+1:\tau_{j+1}}|\gamma)}.$$

In the event that $\tau_j - l \leq \tau_{j-1}$ and $t < \tau_j$, B must be multiplied by $(\tau_j - \tau_{j-1} + l)/(t - \tau_{j-1} + l)$. Similar modifications are needed if $t > \tau_j$ or $\tau_j + l \geq \tau_{j+1}$.

Mixture of updates: A mixture of the two moves above should improve mixing and not be overly computationally expensive. For example, choose the Gibbs update with probability $g_k = 1/\sqrt{k}$ ($k \geq 1$) and random walk with probability $r_k = 1 - g_k$.

5 Prior specification

There are many possible choices for $\pi(z|k, \xi)$. Yao (1984) considers a geometric distribution for the duration, d , of segments; $d \sim \text{Geometric}(p)$. The prior used by Green (1995) has been adapted by Fearnhead (2006) for the discrete time context discussed here. The k changepoint locations are distributed as the even numbered order statistics in a sample of size $2k+1$ from the integers $1, \dots, n-1$, drawn without replacement.

The geometric prior relies on specification of $\xi = p$. Ideally, one could simulate a segment specific p_j in a similar vein to Chib (1998). However this leads to more difficult jump

dynamics when adding or deleting a changepoint. The choice of p may impact the analysis. If too small, then it will assign very small probability to changepoints, meaning small changes cannot be detected with high power. If too large, then spurious changepoints are inferred. For these reasons, it is desirable to introduce a hyperprior on p . For example, a $\text{Beta}(\alpha_1, \alpha_2)$ prior with $1 < \alpha_1 < \alpha_2$ (more weight less than 0.5), would be an ideal choice if there is enough prior information to choose α_1, α_2 . Otherwise, a non-informative $\text{Beta}(1, 1)$ prior would suffice.

Segment parameters share a common hyperparameter γ in Section 2. It is therefore possible to explore uncertainty in γ also by introducing a hyperprior $\pi(\gamma)$.

Sampling p and γ can be easily incorporated into the MCMC scheme in Section 3. One sweep of the algorithm consists of:

1. Sample the changepoints.
2. Conditional on the changepoints sample p .
3. Conditional on the changepoints sample θ .
4. Conditional on θ sample γ and discard the θ values.

For the last step here, it will often be possible to sample γ using a Gibbs step. However, if this is not possible, a simple random walk Metropolis-Hastings could be used.

6 Analysis by filtering recursions

It is useful to give a brief recap of the filtering recursions analysis of Fearnhead (2006) based on a point process prior for changepoint positions. Liu & Lawrence (1999), Barry & Hartigan (1992) have also used these types of methods for the analysis of changepoint problems. Define

$$R_\gamma(t) = \Pr\{y_{t:n} | \text{changepoint at } t-1, \gamma\}.$$

It is possible to compute this quantity in a backward recursion. Defining $R_\gamma(n) = \pi(y_n | \gamma)$, for $t = n-1, \dots, 2$

$$R_\gamma(t) = \sum_{s=t}^n \pi(y_{t:s} | \gamma) R_\gamma(s+1) g(s-t+1) + \pi(y_{t:n} | \gamma) (1 - G(n-t+1))$$

and

$$R_\gamma(1) = \sum_{s=1}^{n-1} \pi(y_{1:s} | \gamma) R_\gamma(s+1) g_0(s) + \pi(y_{1:n} | \gamma) (1 - G_0(n-1))$$

where the dependence of $R_\gamma(t)$ on the hyperparameter γ has been made explicit. Here $g(\cdot)$ gives the point process for the changepoint positions and $G(\cdot)$ the corresponding cumulative distribution function (the subscript 0 on g and G in $R_\gamma(1)$ denotes the distribution of the first changepoint after 0). Yao (1984) takes this as geometric as do Barry & Hartigan (1992). Fearnhead (2006) suggests a negative binomial family in general for this process.

After computing the recursions, a sample of size N of the changepoints can be efficiently simulated as follows:

1. Initialize all samples to have a changepoint at $t = 0$.
2. For $t = 0, \dots, n - 2$
 - (a) Get n_t , the number of samples for which the last changepoint was at time t .
 - (b) If $n_t > 0$ compute the distribution of the next changepoint:

$$\Pr\{\tau|y_{1:n}, t\} = \pi(y_{t+1:\tau}|\gamma)R_\gamma(\tau + 1)g(\tau - t)/R_\gamma(t + 1)$$

- (c) Sample n_t times from $\Pr\{\tau|y_{1:n}, t\}$ and update the n_t samples that have the last changepoint at t .

There are two strengths of this approach. The first is that the samples of changepoints will be independent draws from the posterior distribution. The second is the fast sampling algorithm which avoids computing the distribution of the next changepoint for each possible time. The main weakness of this approach is that the generated samples are dependent on a fixed value of the hyperparameters γ . Updating γ using a hyperprior to correctly explore uncertainty in the value would involve recomputing the recursions $R_\gamma(t)$ for each new value of γ , a computation which is quadratic in n . This would lead to an infeasible computational overhead for any reasonably large sample from the posterior.

7 Poisson data: coal mining disasters

The sampler of Section 4 was applied to the coal-mining data of Jarrett (1979). This data records the dates of serious coal-mining disasters between 1851 and 1962. Disasters are assumed to arise from a Poisson process whose intensity is the height of a step function with an unknown number of steps. For comparison with Fearnhead (2006), time is discretized in weeks and the intensities are taken to be $\text{Gamma}(1, 200/7)$, *a priori*. Details on the model marginal likelihood calculations are given in the Appendix. Conditional on k changepoints the prior on their positions was taken to be the same as the distribution of the even numbered order statistics of a sample of size $2k + 1$ drawn without replacement from $\{1, \dots, n - 1\}$ (Fearnhead 2006),

$$\pi(\tau_1, \dots, \tau_k|k) = \binom{n-1}{2k+1}^{-1} \prod_{j=0}^k (\tau_{j+1} - \tau_j - 1),$$

where for convenience, $\tau_0 = 0$ and $\tau_{k+1} = n$. The algorithm was run for 500,000 sweeps after 10,000 burn in. Every 50th sample was taken to reduce dependency in the MCMC iterates. This took 10 seconds on a 2.5GHz processor. Figure 1 (a) shows that the posterior number of changepoints is almost identical to that obtained from long runs of a RJMCMC sampler and methods based on recursions (see Fearnhead (2006), Figure 1.(a)).

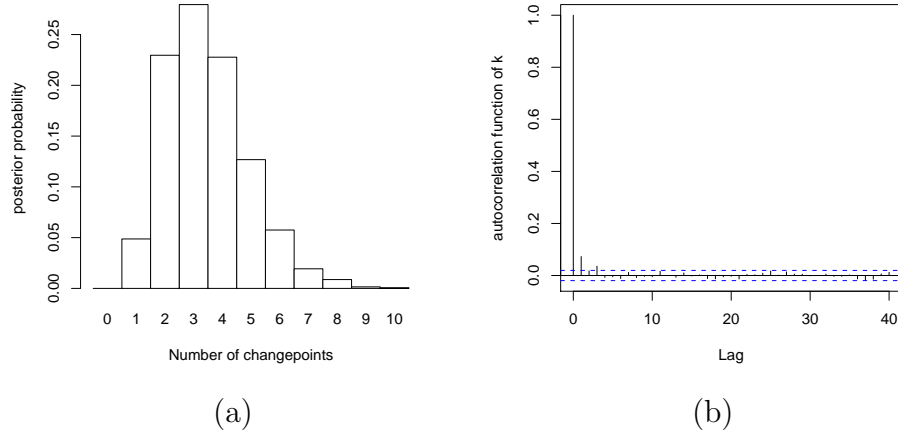


Figure 1: Coal mining disasters: (a) Posterior number of changepoints (b) Plot of the autocorrelation function of the number of changepoints

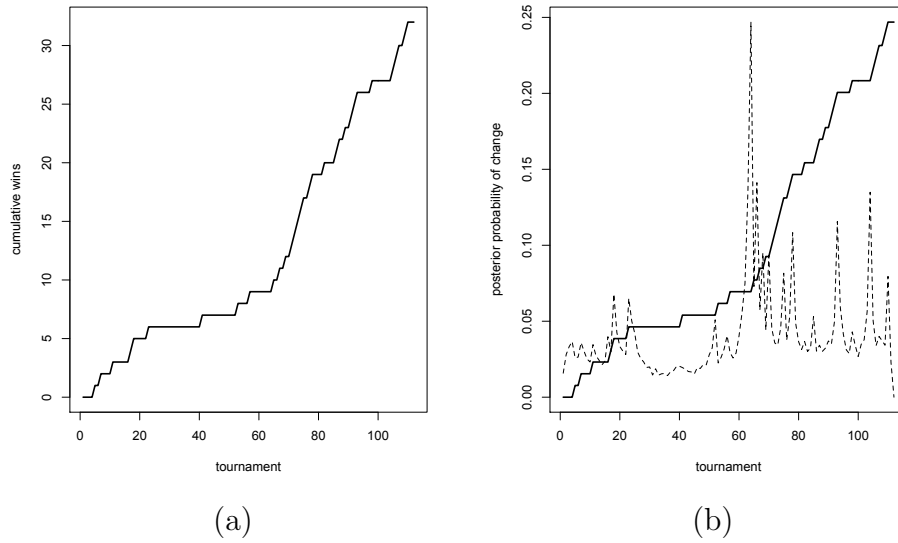


Figure 2: Streakiness dataset: Cumulative counts of Tiger Woods' tournament wins

8 Streakiness in sports

A sportsperson is considered “streaky” if instead of having a constant success rate over time, they have periods of high success rate. Such data will generally be a binary sequence with a “0” denoting a loss and a “1” denoting a win. The data concerning Tiger Woods’ championship wins from September 1996- June 2001 was given and analyzed by Yang (2004), and are reanalyzed using the sampler of Section 4. The cumulative counts are shown in Figure 2 (a). Following Yang (2004) the data as is assumed to arise as a sequence of Bernoulli trials, with a possible changing probability of success. The data is ordered by subsequent tournament, and if a changepoint occurs, it is assumed to do so at some tournament. Let $s_j = \sum_{i=\tau_{j-1}+1}^{\tau_j} y_i$, the number of successes in a segment. Then assuming a Beta(α, β) prior for the probability of success in any segment,

$$\pi(y_{\tau_{j-1}+1:\tau_j} | \alpha, \beta) = \frac{\Gamma\{\alpha + \beta\}}{\Gamma\{\alpha\}\Gamma\{\beta\}} \frac{\Gamma\{s_j + \alpha\}\Gamma\{\tau_j - \tau_{j-1} - s_j + \beta\}}{\Gamma\{\tau_j - \tau_{j-1} + \alpha + \beta\}}.$$

Details of this calculation are given in the Appendix. The parameters α and β were both set equal to 1. The distribution between changepoints was taken to be Geometric(p). The specification of p may have an effect on the outcome of the analysis. It is thus desirable to investigate uncertainty in its value. This is done in two ways. Firstly, a simulation study using the sampler of Section 4 is carried out, where there is a hyperprior placed on p . Secondly, outputs of analyses using filtering recursions (Fearnhead 2006) for a range of values p are compared.

For the MCMC simulation study using the sampler proposed earlier, the hyperparameter given to p was uniform on $[0, 1]$. After each update of the changepoints the value of p was updated by drawing from its full conditional distribution which is Beta($k + 1, n - k$). A discrete uniform prior on $[0, \dots, 10]$ was taken for the number of changepoints. This gives no discriminating prior weight on a particular number of changepoints. The sampler was run 100 times each for 100,000 burn in iterations and a subsequent 1,000,000 iterations. To reduce dependency in the sample, only every 100th sample was stored. Each run took about 1.5 min on a 2.5GHz processor. Changepoints were updated using the mixture of moves discussed in Section 4. Figure 2 (b) shows the output from one of these runs, with the posterior probability of a changepoint at any tournament indicated by the dashed line and a scaled counts curve overlain. Figure 3 (a) shows posterior probability of the number of changepoints over the 100 runs of the sampler. It can be seen that the sampler performs consistently, giving similar results over the 100 runs. Figure 3 (b) shows a histogram for the sampled values of p from the last run. Posterior support for p is highest over the range $[0, 0.1]$.

For the filtering recursions analysis (Fearnhead 2006), the recursions of Section 6 were computed for $p \in [0, 0.1]$ following the analysis above. A sample of size 100,000 changepoints was generated and the posterior of the number of changepoints was computed for each value of p . The modal number of changepoints was recorded from this for each value of p and is shown in Figure 4. It is clear that the number of changepoints inferred in the filtering recursions analysis is very sensitive to the value of p for this data. It is questionable whether such an analysis would be useful for a practitioner since it is unclear how one could objectively choose p in this situation. Certainly an exploratory analysis would be necessary

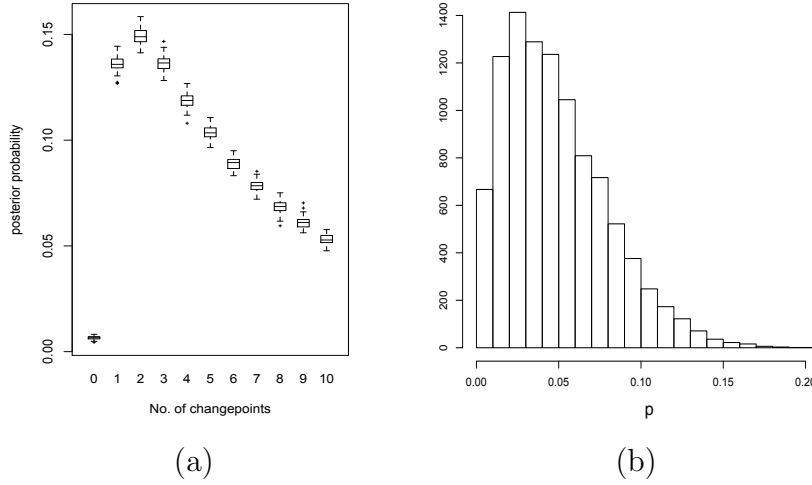


Figure 3: Streakiness dataset: (a) Boxplots of posterior probability for a given number of changepoints for 100 independent runs of the sampler. (b) Histogram of marginal draws of p from one run in the MCMC sampler simulation study.

before choosing the value of p to compute the filtering recursions. One suggestion is to use the sampler proposed here for an exploratory analysis of the posterior allowing for uncertainty in the specification of p . The MCMC sampler simulation study suggests that two changepoints is most likely although there is relatively strong support for up to five changepoints. In this case, specification of one value of p to generate samples of changepoints will not fully explore uncertainty in the posterior. As before, the output of the MCMC sampler shown from Figure 2 (b) shows that one change is clearly identified, but that there is considerable uncertainty in the other positions, hence the support for up to five changepoints.

9 Gaussian changepoint models

Gaussian changepoint models are widely used and studied. Models can include those with changing mean and/or variance across segments. The model assumed for the purposes of the example here is piecewise constant, where data in any segment is Gaussian distributed. Segments share a common error variance. Data point y_i in segment j is assumed to arise independently from a $N(\mu_j, \sigma^2)$ distribution. The segment means μ_j are assumed to arise from a Gaussian distribution with mean μ_0 and variance $\nu^2 \sigma^2$, *a priori*. Denote $\gamma = (\sigma^2, \mu_0, \nu^2)$. Segment length is assumed to have a geometric distribution with parameter p . This gives the log posterior (up to a constant) as

$$\begin{aligned} \log \pi(k, z|y, p, \gamma) = & -(k+1) \log \nu - (n+k+1) \log \sigma + (n-k-1) \log(1-p) + k \log p \\ & - \frac{1}{2} \sum_{j=1}^{k+1} \left\{ \log \left(\tau_j - \tau_{j-1} + \frac{1}{\nu^2} \right) - \frac{1}{\sigma^2} \left(s s_j + \frac{\mu_0^2}{\nu^2} - \frac{(s_j + \frac{\mu_0}{\nu^2})^2}{\tau_j - \tau_{j-1} + \frac{1}{\nu^2}} \right) \right\}, \end{aligned}$$

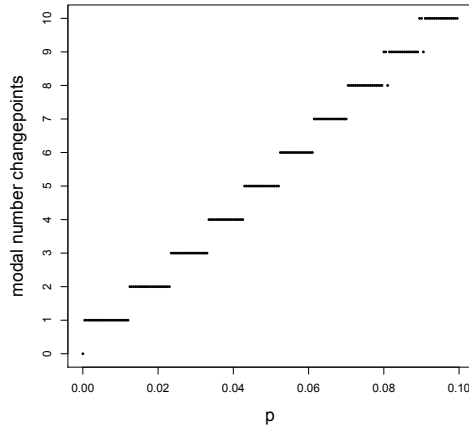


Figure 4: Streakiness data: Modal number of changepoints from a filtering recursions analysis over a range of values of p .

where $ss_j = \sum_{i=\tau_{j-1}+1}^{\tau_j} y_i^2$ and $s_j = \sum_{i=\tau_{j-1}+1}^{\tau_j} y_i$. Details of this calculation are given in the Appendix.

Application to Well-log data

The Well-log data (Ó Ruanaidh & Fitzgerald (1996)) records measurements of nuclear-magnetic response of underground rocks obtained by lowering a probe into a bore-hole. The probe records the response at regular points in time. As well as Fearnhead (2006) this data was also analyzed in Fearnhead & Clifford (2003). The data consists of 4050 measurements, some of which are outliers and were removed before analysis. The data are shown in Figure 5.

The purpose of this example is to demonstrate how results from an analysis with filtering recursions may be sensitive to the choice of hyperparameters γ and how a short run of the sampler could possibly provide good starting values. It is possible to fit a more elaborate state space model to the Well-log data, however, this is not considered here.

Fearnhead (2006) chose the values $p = 0.013, \sigma = 2, 330, \nu = 4.3, \mu_0 = 115,000$ when analyzing the Well-log data in the section on inclusion of hyperpriors. Two simple experiments were performed here to investigate sensitivity of the posterior distribution to prior specification. One of p (Experiment 1) or σ (Experiment 2) was varied over a grid on a small range keeping all other hyperparameter values fixed (details in Table 1). The recursions of Section 6 were computed for each value on the grid and a sample of size 100,000 was generated from the posterior of the changepoints. The empirical posterior distribution of the number of changepoints was computed for each of these samples and the modal number of changepoints recorded. The results are summarized in Figure 6. It can be seen that the modal value of the posterior number of changepoints is sensitive to the values of both p and σ . Thus choosing these values, *a priori*, places the posterior mass $\pi(k, z|y, p, \gamma)$ in the area determined by p and σ and may not correctly represent the true posterior over all p, σ .

Recursion Sensitivity	Fixed	Varied
Experiment 1	$\sigma = 2, 330, \nu = 4.3, \mu_0 = 115, 000$	$p \in [0.005, 0.03]$
Experiment 2	$p = 0.013, \nu = 4.3, \mu_0 = 115, 000$	$\sigma \in [2250, 2750]$

Table 1: Well-log data: Experiments to investigate sensitivity of results of filtering recursions to prior specification

For the Well-log data it would seem most sensible to carry out an analysis with inclusion of hyperpriors on p, σ and μ_0 using the scheme outlined in Section 5. The hyperpriors used are $\pi(p) \propto 1$, $\pi(\mu_0) \propto 1$, $\pi(\nu) \propto 1/\nu$, $\pi(\sigma) \propto 1/\sigma$. The bottom of Figure 5 shows the posterior probability of a change output from an algorithm run for 10,000 burn-in and 100,000 subsequent iterations using a random walk update for changepoint positions. Ergodic mean estimators of the hyperparameters were $\hat{\sigma} = 2360, \hat{p} = 0.014, \hat{\nu} = 3.99, \hat{\mu}_0 = 113771.0$. This took about 10 sec on a 2.5GHz processor with very diffuse starting values. This Gaussian model infers many changepoints as it picks up small changes in the mean and thus performs well for this data.

A long run of the sampler was implemented so as to obtain a near independent sample (1.8×10^7 iterations taking every 1,800th sample; estimated integrated autocorrelation time of the number of changepoints ≈ 1) of size 10,000 from the posterior distribution of changepoints and hyperparameters. This was compared with results from the independence proposal suggested by Fearnhead (2006). In the independence proposal MCMC scheme suggested in Fearnhead (2006), a sample of changepoints is generated using filtering recursions conditional on $p = 0.013, \sigma = 2, 330, \mu_0 = 115, 000, \nu = 4.3$. This sample is then used for an independence proposal and hyperparameters are updated in the same way as done here. Figure 7 shows kernel density estimates constructed from samples of the hyperparameters for the sampler (dashed line) and independence proposal (solid line). It can be seen that there is a slight discrepancy in that the independence proposal leads to more peaked densities.

In our implementation an independence proposal based on a sample of size 10,000 was used. This updating scheme for hyperparameters and changepoints was then run for 50,000 iterations. Although the acceptance rate for moving between different changepoint configurations was high, the independence proposal distribution was highly degenerate. Only ten unique changepoint configurations were sampled in the 50,000 iterations of the MCMC scheme. For other datasets where less information is available to choose the hyperparameters to generate the independence proposal, it is possible that this could lead to highly biased sampling from the hyperpriors.

In the sense of hyperprior incorporation and full exploration of the posterior distribution the MCMC sampler proposed performs better than the independence proposal. However, generating independent samples may be more costly in large datasets with many changepoints. Nonetheless, it is clear that the inclusion of hyperpriors circumvents the sensitivity of posterior distribution of the changepoints to specification of the hyperparameters. This is a main advantage of the approach proposed here and makes the detection of changepoints more automatic.

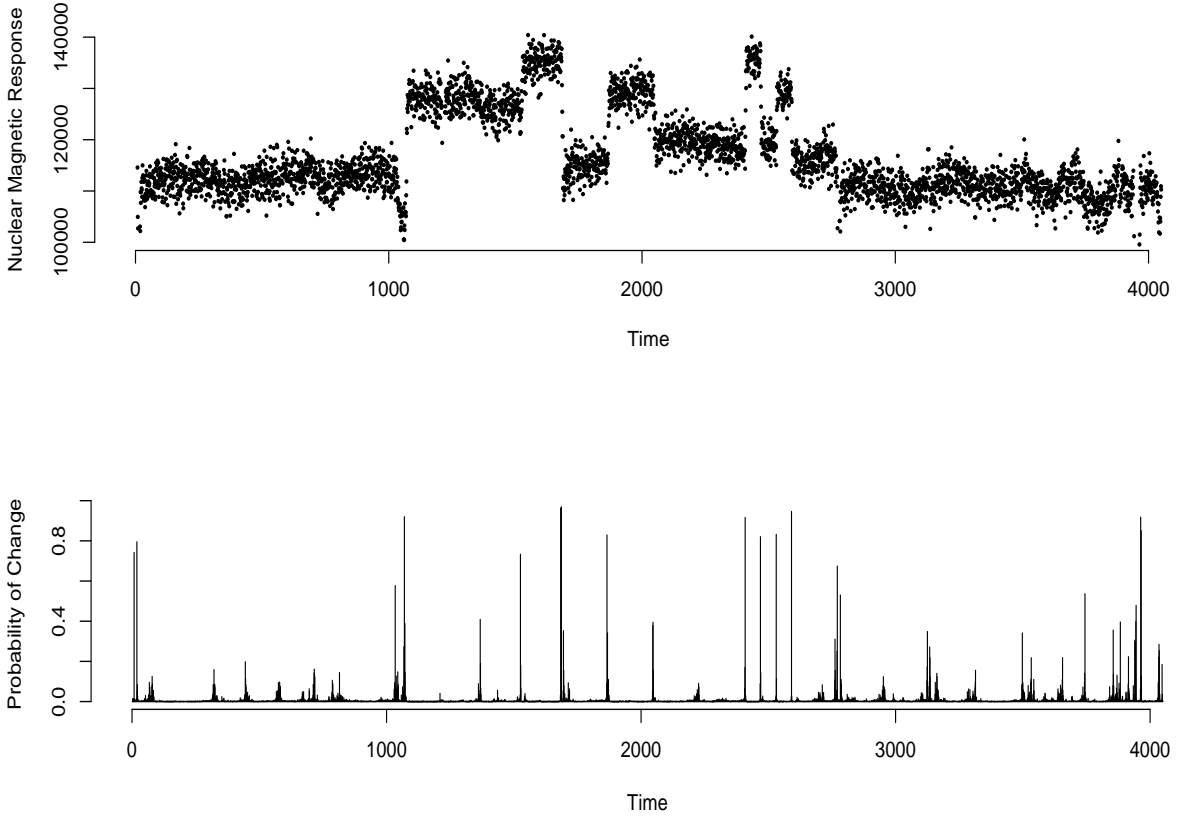


Figure 5: Top: Well-log data. Bottom: Posterior probability of a changepoint in any position from 100,000 samples using the sampler with hyperpriors.

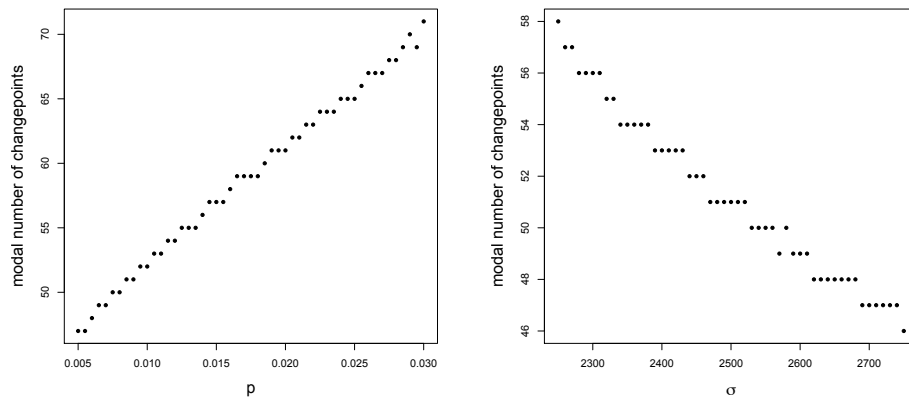


Figure 6: Well-log data: Modal number of changepoints for a filtering recursions analysis of the Well-log data for Experiment 1 and Experiment 2. Experiment 1 varies p (left) and Experiment 2 varies σ (right)

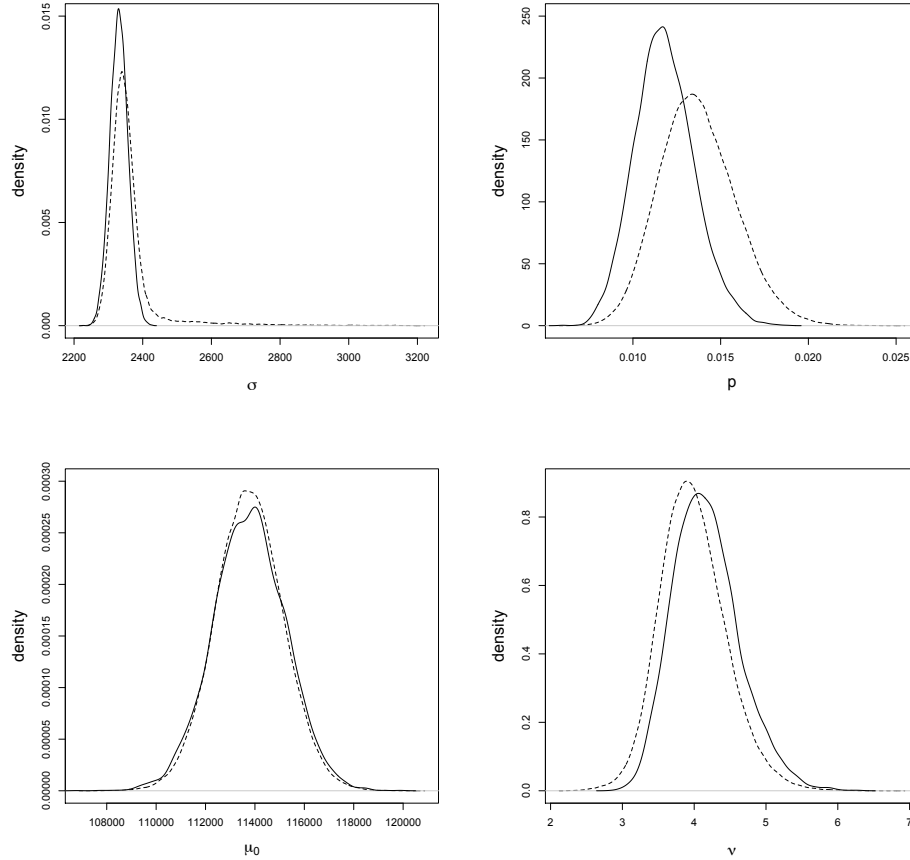


Figure 7: Well-log data: Comparison of long run of sampler to MCMC scheme with independent proposals from filtering recursions. Dashed lines give the density from the MCMC sampler output and solid lines give the density output from analysis using the independent proposal scheme suggested in Fearnhead (2006).

10 Discussion

This paper has presented an MCMC method to perform retrospective inference for changepoint model which are collapsible. The multiple changepoint problem is rephrased as a stochastic model search over a large models space, with the Bayes factors for competing models appearing in the acceptance probabilities for the MCMC sampling scheme.

The performance of the sampler was verified for the benchmark coal mining disasters data. Application of the sampler to a streakiness dataset from sports revealed that posteriors for the number of changepoints can be diffuse. It was demonstrated that prior specification on the duration of segments plays a crucial role in the analysis of the models considered. Incorporation of hyperpriors to account for this revealed features of the posterior that would be missed by a popular filtering recursions analysis for changepoints. Application to the Well-log data further highlighted sensitivity of analysis by filtering recursions to prior specification. It was shown that output from a short run of our sampler can be used to give good values of the hyperparameters for this prior specification.

In conclusion, the sampling scheme presented is shown to work well and can provide further insight and account for prior uncertainty in some difficult situations. It can be used as a useful exploratory tool or for a full analysis. Computer code implementing the sampler written in C may be downloaded from www.ucd.ie/statdept/jwyse.

Appendix

Calculations for the coal-mining example

Given a segment $y_{s:t}$, each $y_i \sim_{\text{iid}} \text{Poisson}(\mu)$. Here μ is the height of the step function that gives the intensity of the process between times s and t . Assume the prior for μ is $\text{Gamma}(\rho, \lambda)$ where $\gamma = (\rho, \lambda)$. The marginal likelihood for the segment is then

$$\begin{aligned} \pi(y_{s:t}|\gamma) &= \int_0^\infty \frac{\lambda^\rho}{\Gamma\{\rho\}} \mu^{\rho-1} \exp\{-\lambda\mu\} \prod_{i=s}^t \frac{\mu^{y_i}}{y_i!} \exp\{-\mu\} d\mu \\ &= \frac{\lambda^\rho}{\Gamma\{\rho\}} \int_0^\infty \frac{1}{F_{s:t}} \mu^{S_{s:t}+\rho-1} \exp\{-(t-s+\lambda+1)\mu\} d\mu \end{aligned}$$

where $F_{s:t} = \prod_{i=s}^t y_i!$ and $S_{s:t} = \sum_{i=s}^t y_i$. Completing the integral of the Gamma density gives

$$\pi(y_{s:t}|\gamma) = \frac{\lambda^\rho}{\Gamma\{\rho\}} \frac{1}{F_{s:t}} \frac{\Gamma\{S_{s:t} + \rho\}}{(t-s+\lambda+1)^{S_{s:t}+\rho}}$$

Calculations for the streakiness example

Within a segment $y_{s:t}$, $y_i \sim_{\text{iid}} \text{Bernoulli}(\phi)$. Taking a $\text{Beta}(\alpha, \beta)$ prior on ϕ , the marginal likelihood is obtained from

$$\pi(y_{s:t}|\gamma) = \int_0^1 \frac{\Gamma\{\alpha + \beta\}}{\Gamma\{\alpha\}\Gamma\{\beta\}} \phi^{\alpha-1} (1-\phi)^{\beta-1} \prod_{i=s}^t \phi^{y_i} (1-\phi)^{1-y_i} d\phi$$

where $\gamma = (\alpha, \beta)$. This reduces to

$$\pi(y_{s:t}|\gamma) = \frac{\Gamma\{\alpha + \beta\}}{\Gamma\{\alpha\}\Gamma\{\beta\}} \int_0^1 \phi^{S_{s:t}+\alpha-1} (1-\phi)^{t-s-S_{s:t}+\beta} d\phi.$$

where $S_{s:t} = \sum_{i=s}^t y_i$. Completing the Beta integral gives

$$\pi(y_{s:t}|\gamma) = \frac{\Gamma\{\alpha + \beta\}}{\Gamma\{\alpha\}\Gamma\{\beta\}} \frac{\Gamma\{S_{s:t} + \alpha\}\Gamma\{t - s + 1 - S_{s:t} + \beta\}}{\Gamma\{t - s + 1 + \alpha + \beta\}}.$$

Calculations for Gaussian changepoint model

The model for all the data may be written hierarchically as

$$\begin{aligned} \pi(k, z, \theta|y, p, \gamma) &\propto \pi(z|k, p)\pi(\theta|k, z, \sigma, \mu_0)\pi(y|k, z, \theta) \\ &\propto p^k(1-p)^{n-k-1} \prod_{j=1}^{k+1} \frac{1}{\nu\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\nu^2\sigma^2}(\mu_j - \mu_0)^2\right\} \\ &\quad \times \prod_{i=\tau_{j-1}+1}^{\tau_j} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_j)^2\right\} \\ &= \frac{(2\pi)^{-(n+k+1)/2}}{\nu^{k+1}\sigma^{n+k+1}} p^k(1-p)^{n-k-1} \\ &\quad \prod_{j=1}^{k+1} \exp\left\{-\frac{1}{2\sigma^2} \left[\left(\tau_j - \tau_{j-1} + \frac{1}{\nu^2} \right) \mu_j^2 - 2 \left(s_j + \frac{\mu_0}{\nu^2} \right) \mu_j + s s_j + \frac{\mu_0^2}{\nu^2} \right] \right\}. \end{aligned}$$

Completing the square on μ_j and then performing integration of μ_j over $(-\infty, \infty)$ gives the required posterior.

$$\begin{aligned} \pi(k, z, \theta|y, p, \gamma) &\propto \frac{(2\pi)^{-n/2}}{\nu^{k+1}\sigma^n} p^k(1-p)^{n-k-1} \\ &\quad \prod_{j=1}^{k+1} \left(\tau_j - \tau_{j-1} + \frac{1}{\nu^2} \right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2} \left(s s_j + \frac{\mu_0^2}{\nu^2} - \frac{(s_j + \frac{\mu_0}{\nu^2})^2}{\tau_j - \tau_{j-1} + \frac{1}{\nu^2}} \right) \right\} \end{aligned}$$

References

- Barry, D. & Hartigan, J. A. (1992), ‘Product Partition Models for Change Point Problems’, *The Annals of Statistics* **20**, 260–279.
- Chib, S. (1998), ‘Estimation and comparison of multiple change-point models’, *Journal of Econometrics* **86**, 221–241.
- Fearnhead, P. (2006), ‘Exact and efficient Bayesian inference for multiple changepoint problems’, *Statistics and Computing* **16**, 203–213.

- Fearnhead, P. & Clifford, P. (2003), ‘On-Line Inference for Hidden Markov Models via Particle Filters’, *Journal of the Royal Statistical Society, Series B* **65**, 887–899.
- Fearnhead, P. & Liu, Z. (2007), ‘On-line inference for multiple changepoint problems’, *Journal of the Royal Statistical Society, Series B* **69**, 589–605.
- Girón, F. J., Moreno, E. & Casella, G. (2007), Objective Bayesian Analysis of Multiple Changepoints for Linear Models, *in* ‘Bayesian Statistics 8’, Oxford University Press, pp. 227–252.
- Green, P. (1995), ‘Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model determination’, *Biometrika* **82**, 711–732.
- Jarrett, R. G. (1979), ‘A note on the intervals between coal-mining disasters’, *Biometrika* **66**, 191–193.
- Lavielle, M. & Lebarbier, E. (2001), ‘An application of MCMC methods for the multiple change-points problem’, *Signal Processing* **81**, 39–53.
- Liu, J. S. & Lawrence, C. E. (1999), ‘Bayesian inference on biopolymer models’, *Bioinformatics* **15**, 38–52.
- Ó Ruanaidh, J. J. K. & Fitzgerald, W. J. (1996), *Numerical Bayesian Methods applied to Signal Processing*, Springer, New York.
- Punskaya, E., Andrieu, C., Doucet, A. & Fitzgerald, W. J. (2002), ‘Bayesian Curve Fitting Using MCMC With Applications to Signal Segmentation’, *IEEE Transactions on Signal Processing* **50**, 747–757.
- Raftery, A. E. & Akman, V. E. (1986), ‘Bayesian Analysis of a Poisson Process with a Change-Point’, *Biometrika* **73**, 85–89.
- Yang, T. Y. (2004), ‘Bayesian binary segmentation procedure for detecting streakiness in sports’, *Journal of the Royal Statistical Society, Series A* **167**, 627–637.
- Yao, Y.-C. (1984), ‘Estimation of a Noisy Discrete-Time Step Function: Bayes and Empirical Bayes Approaches’, *The Annals of Statistics* **12**, 1434–1447.